Challenges and issues relating to the use of Representation Information in the digital curation of Crystallography and Engineering data

3rd International Digital Curation Conference
"Curating our Digital Scientific Heritage: a Global Collaborative Challenge"
12-13th December 2007
Washington DC, USA

Manjula Patel and Alexander Ball
UKOLN, University of Bath, UK

Funded by:
JISC
e-Science

Overview

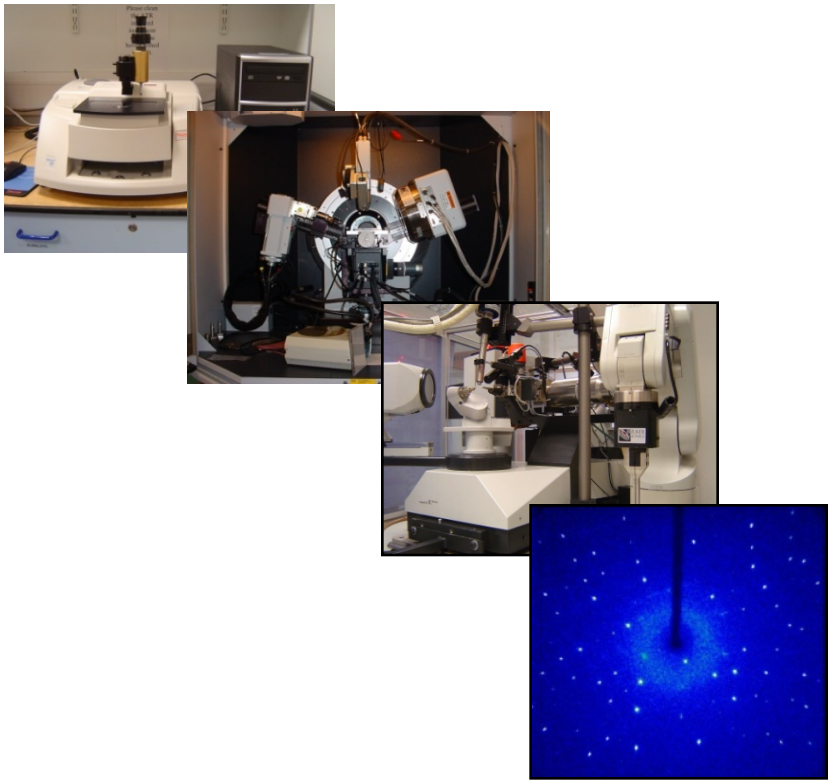
- eBank-UK Project (Crystallography)
- Knowledge & Information Management Project (Engineering)
- OAIS and Representation Information
- Registry/Repository of Representation Information (RRoRI)
- Capturing Representation Information
  - Crystallographic Information File (CIF)
  - Initial Graphics Exchange Specification (IGES) 5.3
- Challenges and Issues
- Concluding Comments

eBank-UK Project (Crystallography)

- Phenomenal growth in amount of data generated from experiments
- Only a small proportion is widely and easily accessible
- eCrystals data repository: rapid dissemination derived and results data from crystallography experiments
- Linking research data to publications and scholarly communication
- JISC funded; three phases Sept. 2003-June 2007
- eBank-UK Phase 3: "A Study of Curation and Preservation issues in the eCrystals Data Repository and proposed Federation", Sept. 2007
  - audit and certification (TRAC, DRAMBORA, NESTOR, ISO International repository audit and certification BOF Group)
  - OAIS and Representation Information for crystallography data
  - eBank-UK application profile and preservation metadata
  - e-Prints.org repository platform

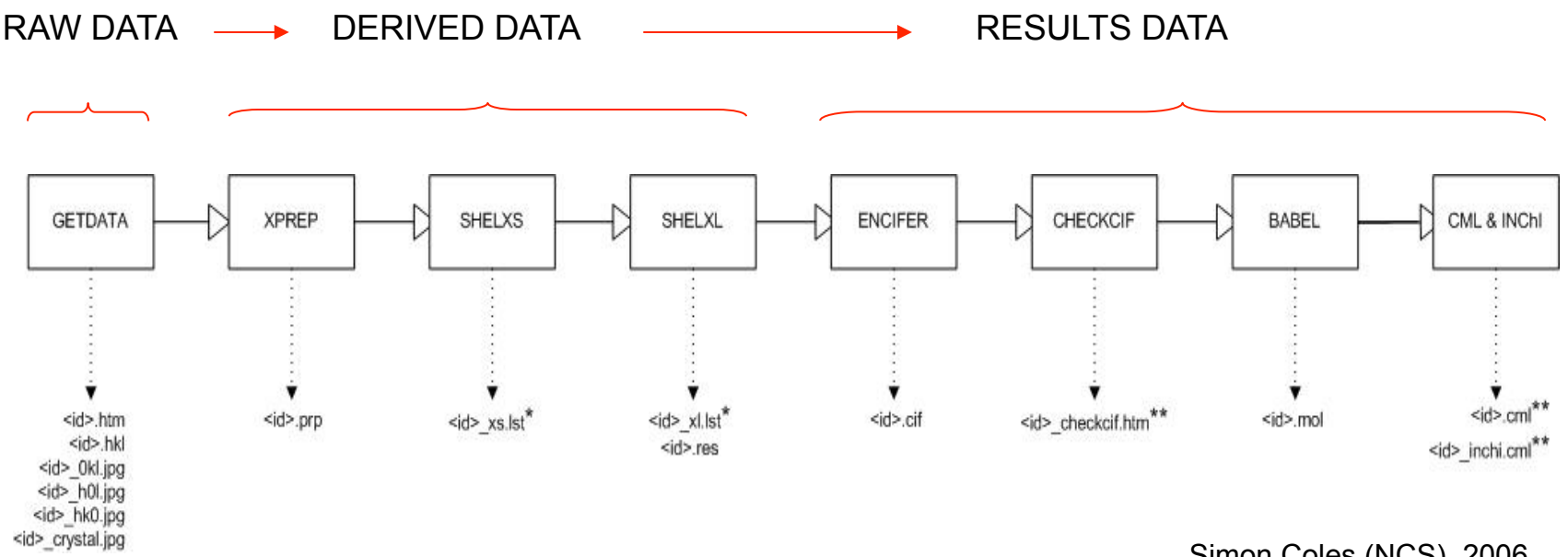
Crystallography –The Science

- Sub-discipline of chemistry
- Concerned with determining the structure of a molecule and its 3D orientation with respect to other molecules in a crystal
- Analysis of diffraction patterns obtained from X-ray scattering experiments
- eBank-UK focused on laboratory based experimental technique of chemical crystallography undertaken at the UK National Crystallography Service (NCS)



Simon Coles (NCS), 2006

Crystal Structure Determination Workflow



Simon Coles (NCS), 2006

- Initialisation: mount new sample, set up data collection
- Collection: collect data
- Processing: process and correct images
- Solution: solve structures
- Refinement: refine structure
- CIF: produce Crystallographic Information File
- Validation: chemical & crystallographic checks
- Report: generate Crystal Structure Report
- CML, INChI

eCrystals: Example Crystal Structure Report

University of Southampton Crystal Structure Report Archive

Home
About
Browse
User Area
Help

2,2-trimethylenedioxy-4,4,6,6-tetrachlorocyclotriphosphazene

Sample Originator: D.B. Davies[a], R.A. Shaw[a], A. Kilic[b], M. Odlyha[a] and A. Uslu[b].

Data Collection: S.J. Coles[c], L.S. Huth[c] and M.E. Light[c].

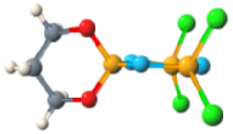
Structure Determination: S.J. Coles[c], J.S. Rutherford and M.B. Hursthouse.

Birkbeck College[a]
Gebze Institute of Technology[b]
University of Southampton[c]

$C_3H_6Cl_4N_3O_2P_3$

InChI=1/C3H12Cl4N3O2P3/c4-13(5)8-14(6,7)10-15(9-13)11-2-1-3-12-15/h8-10,13-15H,1-3H2

Compound Class: Inorganic
Keywords: cyclophosphazene, phase transition, variable temperature
Creation Date: 28 March 2007
Deposited By: Dr Simon J Coles
Deposited On: 28 March 2007



Available Files

Final Result

2005sjc0007.cif 11k
2005sjc0007.cml 4k

Validation

2005sjc0007_checkcif.htm 9k

Data collection parameters

Data collection parameters

Chemical formula	C3 H6 Cl4 N3 O2 P3
Crystallisation Solvent	
Crystal morphology	Rod
Crystal system	Orthorhombic
Space group symbol	Pna2(1)
Cell length a	13.4804(14)
Cell length b	10.6442(9)
Cell length c	8.8479(7)
Cell angle alpha	90.00
Cell angle beta	90.00
Cell angle gamma	90.00
Data collection temperature	274(2)

Refinement results

Solution figure of merit	0.0569
R Factor (Obs)	0.0334
R Factor (All)	0.0380
Weighted R Factor (Obs)	0.0871
Weighted R Factor (All)	0.0905

Citation: D.B. Davies, L.S. Huth, M.B. Hursthouse, M. Odlyha, S.J. Coles, R.A. Shaw, J.S. Rutherford, A. Kilic, M.E. Light, A. Uslu (2007). Southampton, UK, University of Southampton, Crystal Structure Report Archive. (doi:)

2005sjc0007_checkcif.htm 9k

Refinement

2005sjc0007.res 5k
2005sjc0007_xl.lst 29k

Solution

2005sjc0007.prp 5k
2005sjc0007_xs.lst 44k

Processing

2005sjc0007.hkl 532k
2005sjc0007.htm 11k
2005sjc0007_0kl.jpg 91k
2005sjc0007_h0l.jpg 87k
2005sjc0007_hk0.jpg 79k

Data Collection

2005sjc0007_crystal.jpg 17k

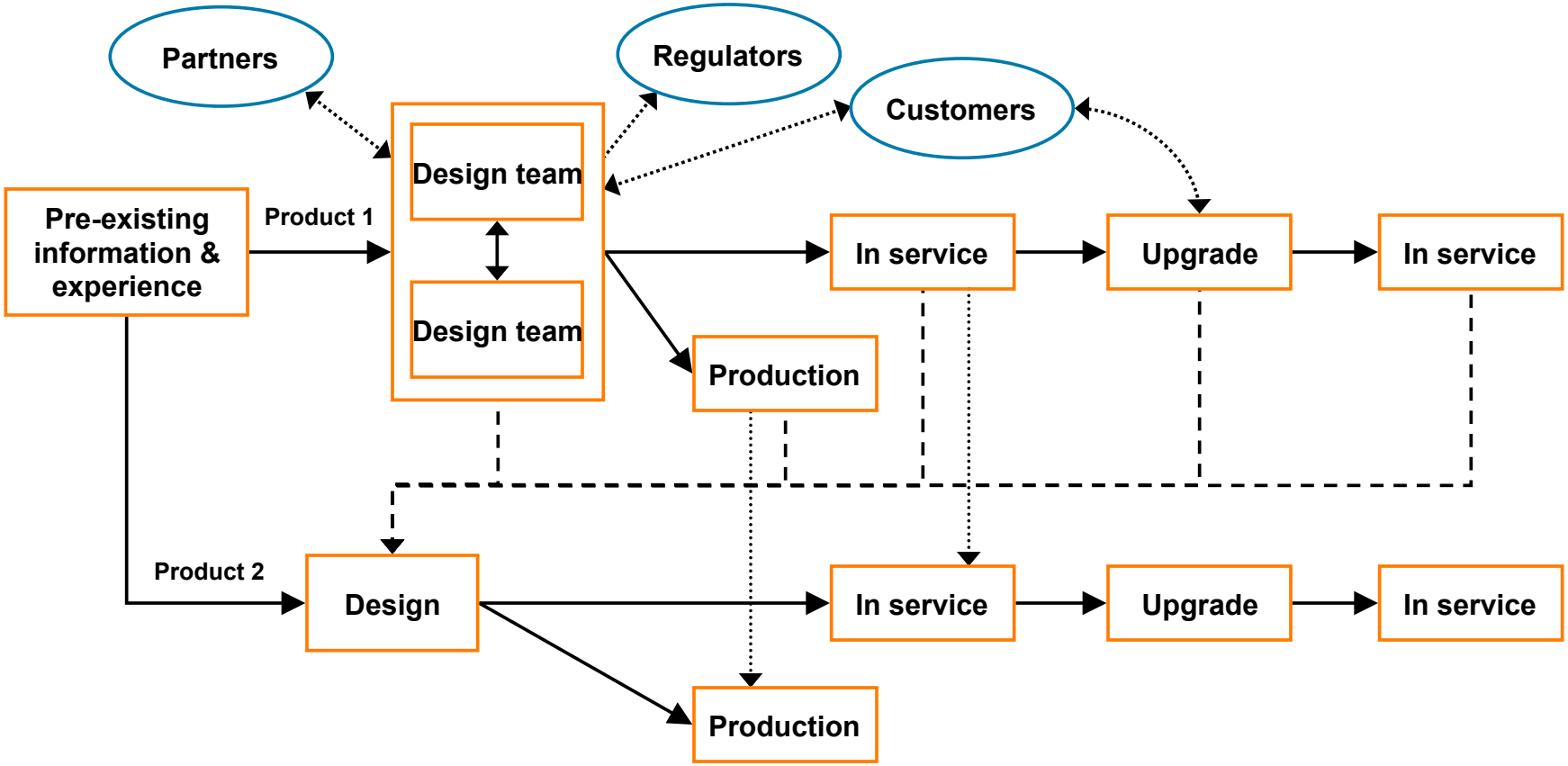
Other Files

2005sjc0007.doc 186k
2005sjc0007.fcf 138k

Knowledge & Information Management through Life Project (Engineering)

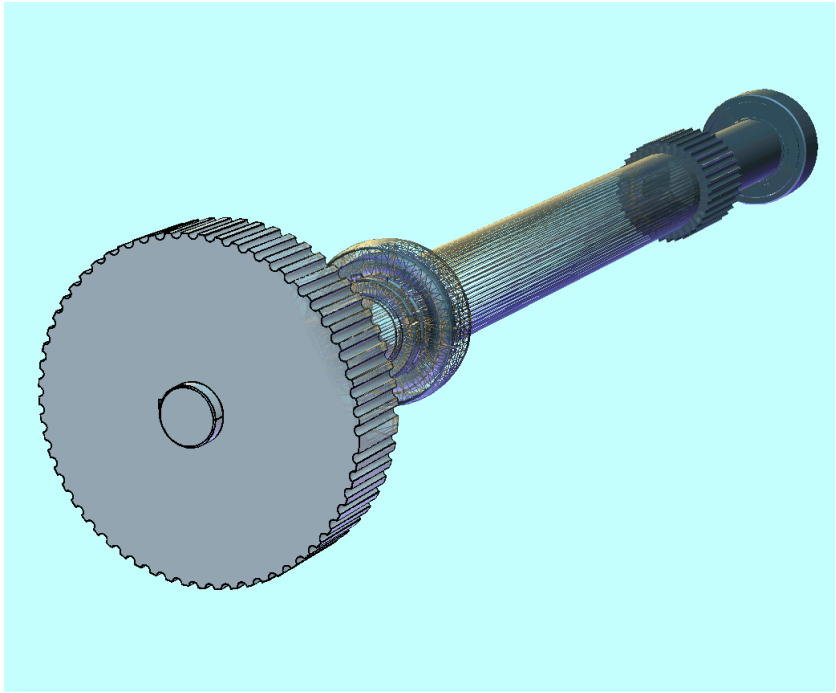
- Switch from product-delivery to product-service paradigm
- Develop tools and techniques for sustainable representation of product, process and design rationale
- Develop approaches to learning about products in service – the performance of the artefact and its impact on users
- Investigate the organisational challenges associated with managing the whole life-cycle of complex product-service systems
- Develop an intellectual framework for the above
- 11 Academic partners
- Industrial partners: construction; aerospace, defence suppliers; MOD; NHS
- £5.5 million total funding, £3.94 million UK EPSRC/ESRC
- Duration: Oct 2005-Mar 2009

Engineering information flows

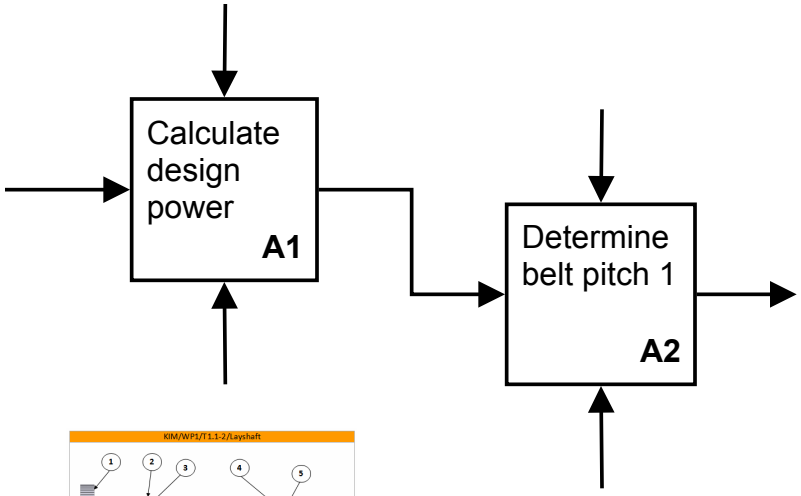


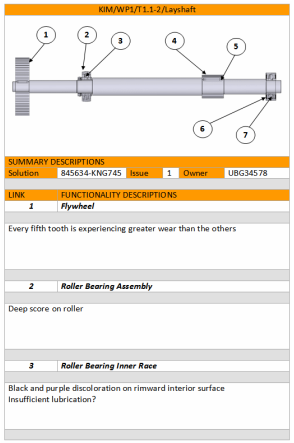
Engineering data objects (1)

- CAD models
  - Geometry
  - Dimensions
  - Tolerances
  - Materials, finishes
  - Feature semantics
  - Model history
- Analytical models
  - Finite element analysis
  - Stress/load bearing



Engineering data objects (2)




KIM/WP1/T1.1-2/Layshaft

SUMMARY DESCRIPTIONS

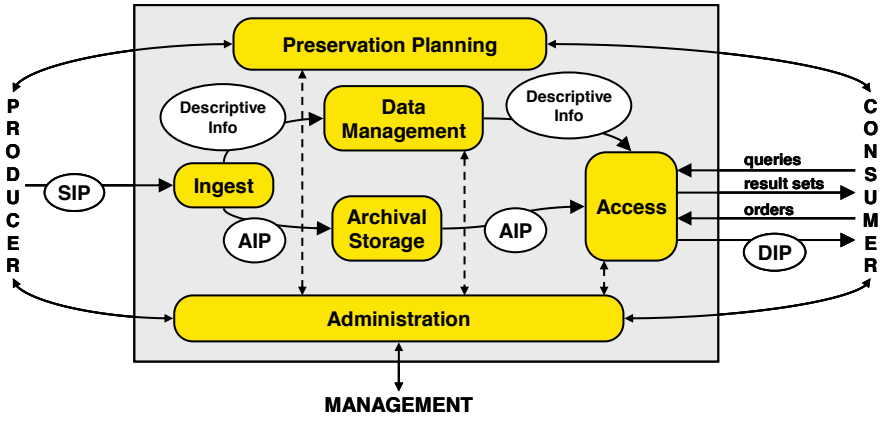
Solution	845634-KNG745	Issue	1	Owner	UBG34578

LINK	FUNCTIONALITY DESCRIPTIONS
1	*Flywheel*
	Every fifth tooth is experiencing greater wear than the others
2	*Roller Bearing Assembly*
	Deep score on roller
3	*Roller Bearing Inner Race*
	Black and purple discoloration on rimward interior surface Insufficient lubrication?

- Design process models
- Manufacturing process models
- Numerical control programmes
- Parts catalogues
- Design reports
- Incident books
- Service record sheets
- . . .

OAIS Functional Entities

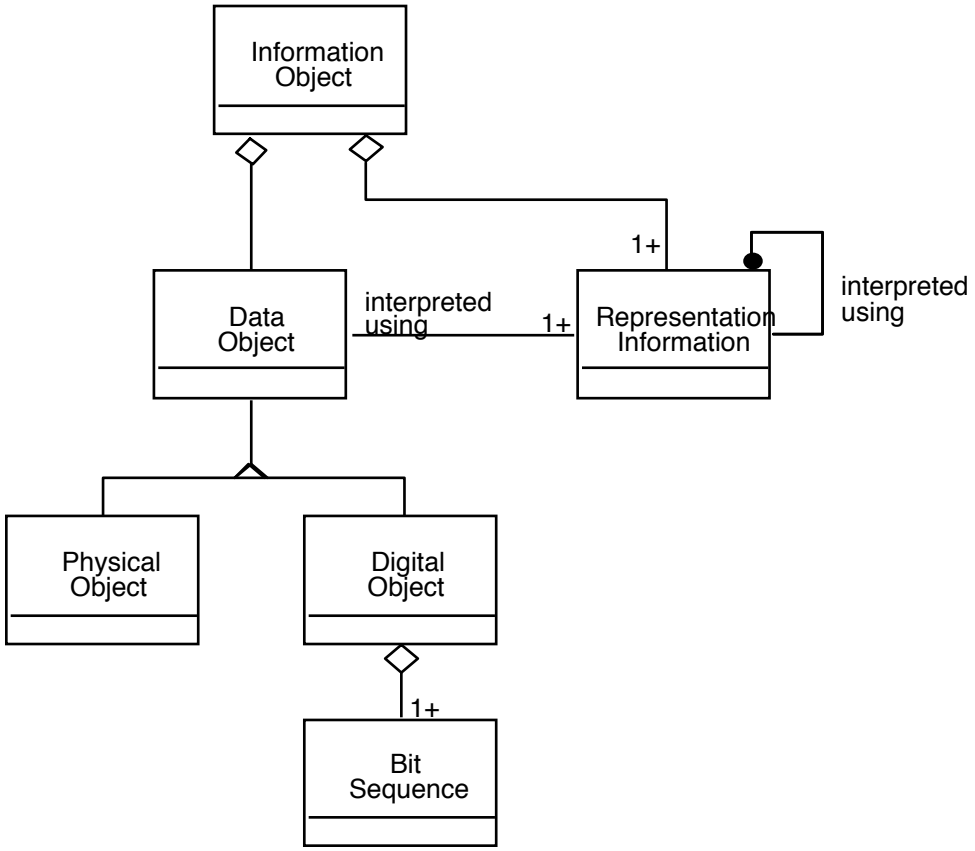
- Ingest: services and functions that accept SIPs from Producers; prepares AIPs for storage, and ensures that AIPs and their supporting Descriptive Information become established within the OAIS
- Archival Storage: services and functions used for the storage and retrieval of AIPs
- Data Management: services and functions for populating, maintaining, and accessing a wide variety of information
- Administration: services and functions needed to control the operation of the other OAIS functional entities on a day-to-day basis



- Preservation Planning: services and functions for monitoring the OAIS environment and ensuring that content remains accessible to the Designated Community
- Access: services and functions which make the archival information holdings and related services visible to Consumers

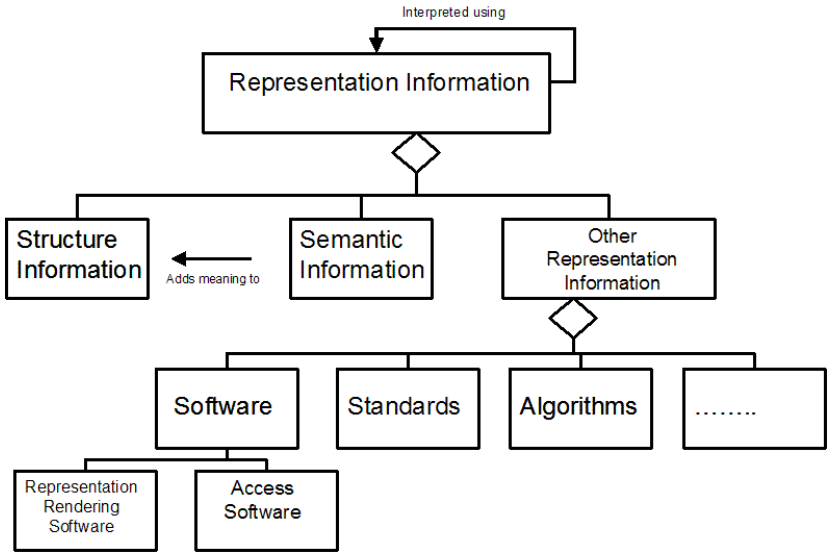
OAIS Information Model

- Information Object is composed of a Data Object that is either physical or digital, as well as the Representation Information that allows for the full interpretation of the data into meaningful information
- Representation Information is *any* information required to render, interpret and understand data



OAIS Representation Information (RI)

- Types of RI
  - Structure
    - e.g. file formats for text, images, audio, moving images, datasets, 3D models
  - Semantic
    - e.g. data dictionaries and knowledge organisation systems such as schemata, ontology, metadata vocabularies and thesauri
  - Other
    - e.g. software, algorithms, standards, time dependent information, actions, processes
- RI is recursive in nature; using one element of RI in a meaningful manner may well require further RI, resulting in a RI Network



- Recursion is terminated based on the designated community's knowledge base
- Essential that RI itself is curated and preserved to maintain access to data (render, interpret and understand)

Registry/Repository of RI (RRoRI)

- Development started under the DCC-Development team
- Work now being undertaken jointly with the CASPAR Project
  - Cultural, Artistic and Scientific knowledge for Preservation, Access and Retrieval (Integrated Project co-funded by EU FP6 Programme, April 2006)
- Representation Information is the key to long-term access
- RRoRI should itself be a trustworthy OAIS
- Repository: some RI is stored; Registry: links to external RI
- Emphasis on interoperability and automated use
- Vision is to have a global, distributed network of RI
- Provide an infrastructure of reliable and trusted RI for third party use

RRoRI: Curation Persistent Identifier

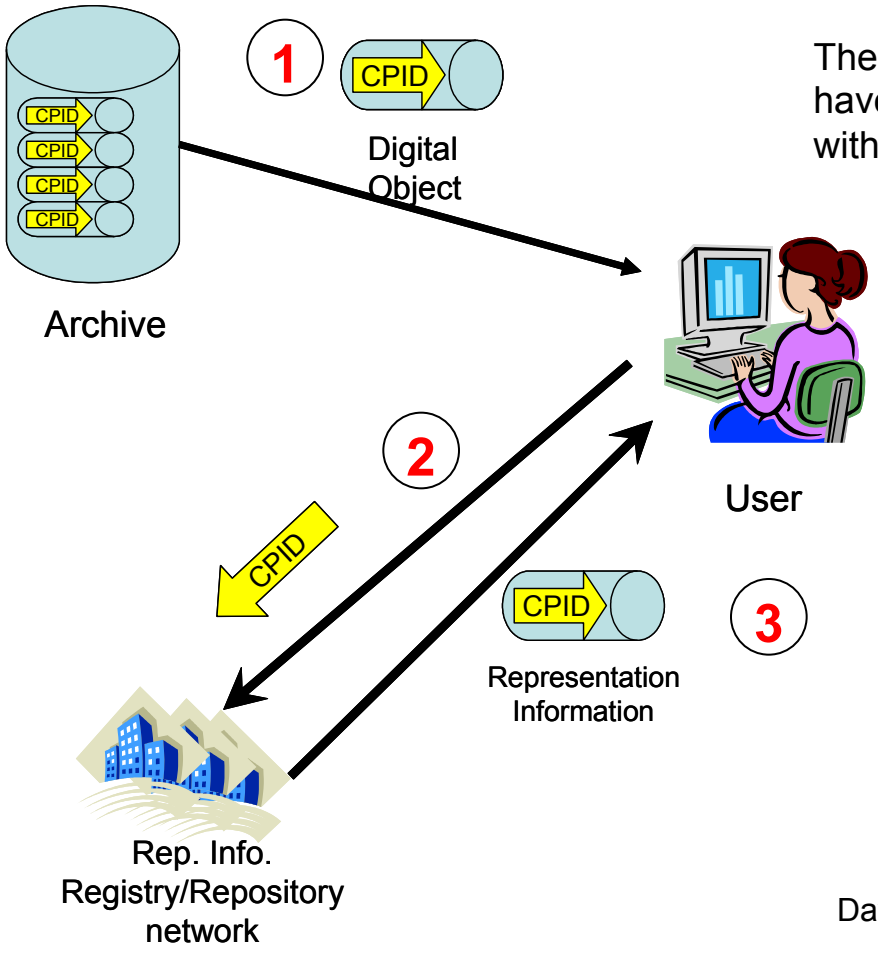
- Idea of RI is the key
  - Information Object: a specific object to be archived/preserved/ curated
  - RI: all information required to render, interpret and understand the object
  - RI Label: used to connect RI to an Information Object
- RI Label serves as a mechanism for accessing RI in RRoRI
  - Label is used to identify relevant RI
  - Provides mechanism for recording individual RI components
- RI Label has a Curation Persistent Identifier (CPID)
  - Used to connect the digital object to the RI Label

Use of CPID

1 User gets data from archive. Data has associated Curation Persistent Identifier (CPID)

2 User unfamiliar with data so requests RI using CPID

3 User receives RI – which has its own CPID in case it is not immediately usable



The Digital Object could have some RI packed with it, as well as a CPID

CPID supports automated access & processing

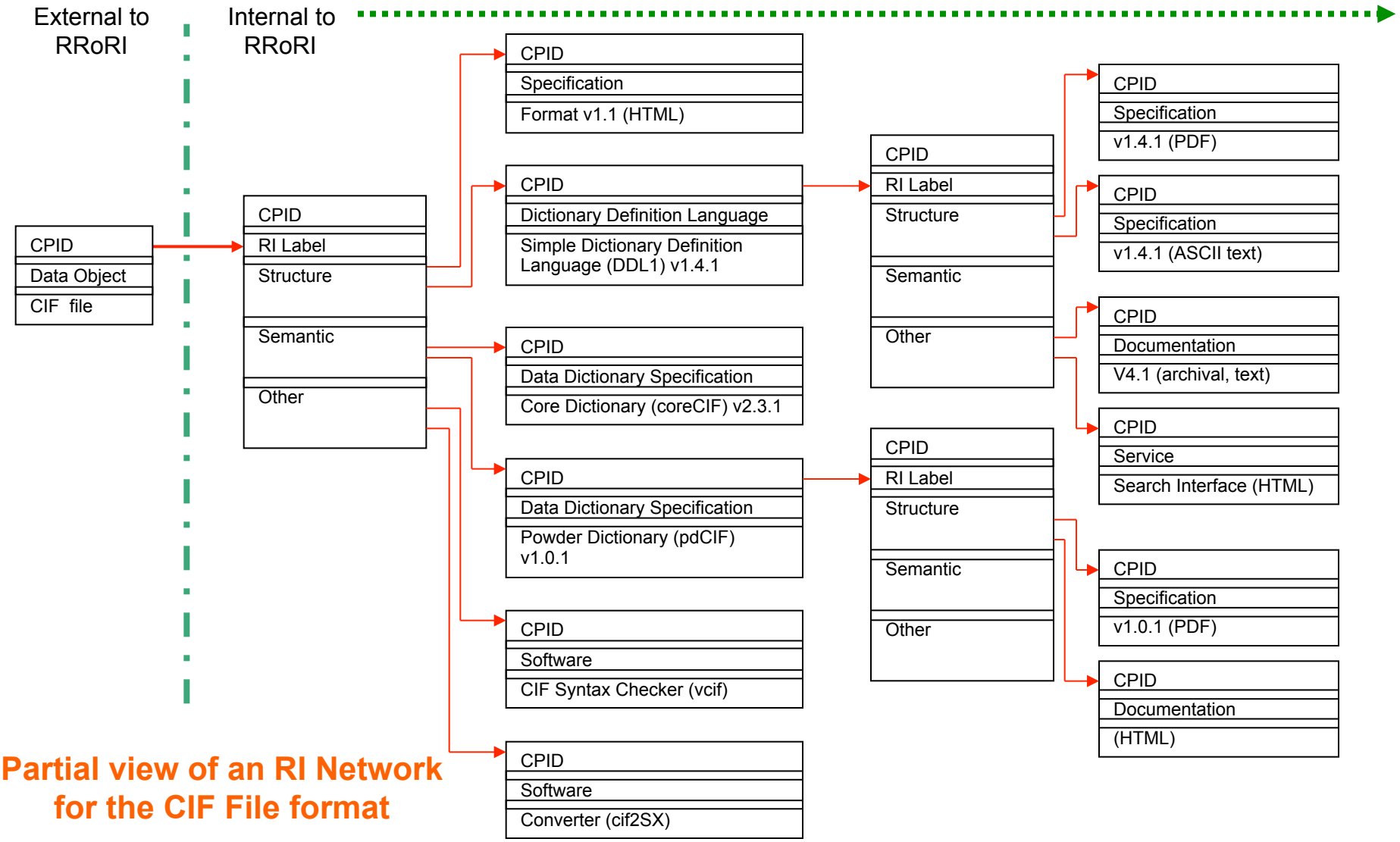
David Giaretta (STFC), 2007

RRoRI: Current RI Classification

- Structure
  - Formats
    - Descriptive Language Specification
    - Digital File Type
    - Specification
- Semantic
  - Data
    - Dictionary Specification
      - Dictionary
  - Document
  - Language
    - Computer Programming Language
    - Human Written Language
  - Models
  - Standards
    - Developing Organisation
- Other
  - Access software
  - Algorithms
  - Computer hardware
    - BIOS
    - CPU
    - Graphics
    - Hard Disk Controller
    - Interface
    - Network
  - Media
  - Physical
  - Processing software
  - Representation Rendering software

Capturing RI: Crystallography Data

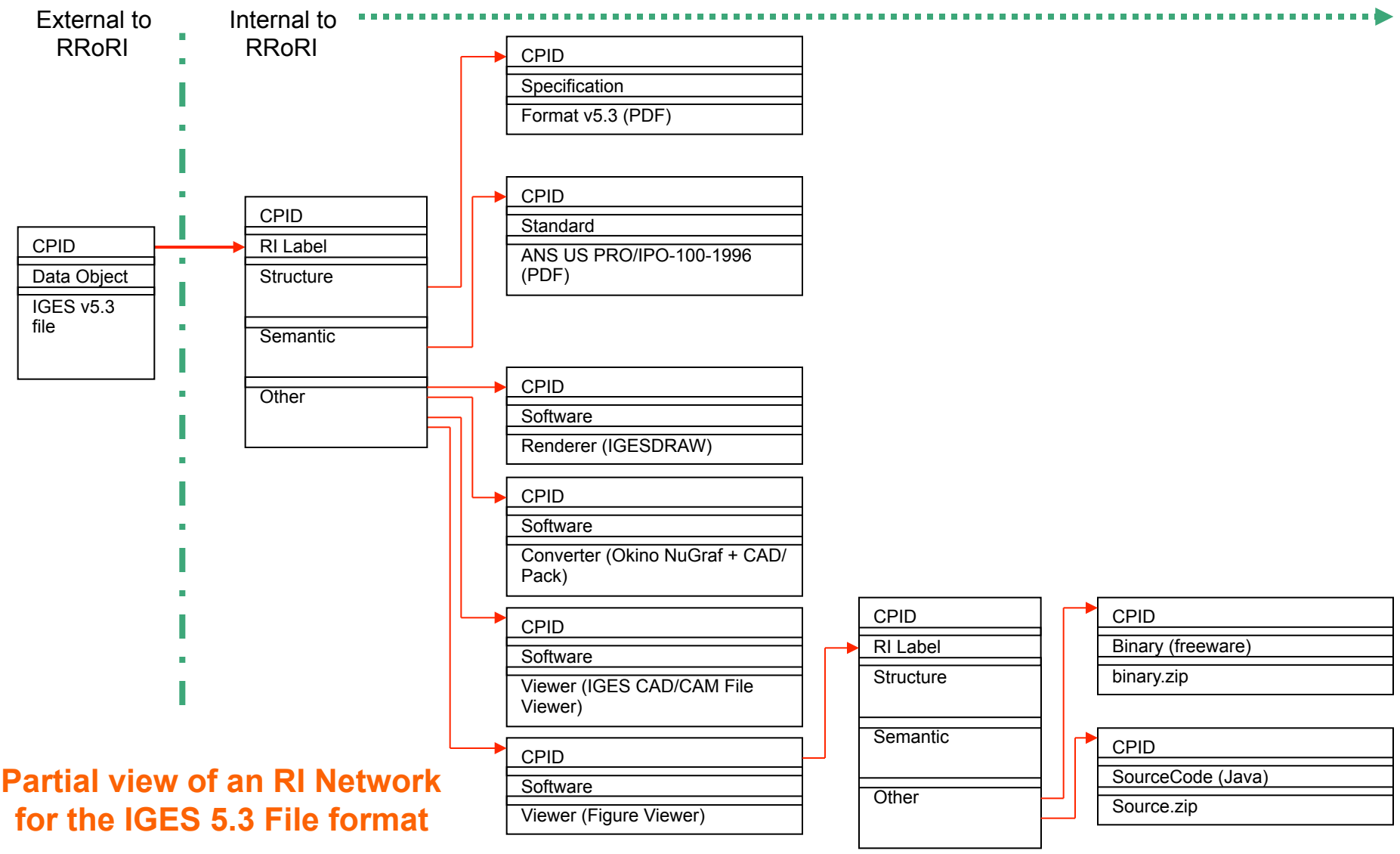
- Bounded domain (within an academic environment)
- Limited number of major stakeholders
  - International Union of Crystallography (IUCr)
  - UK National Crystallography Service (NCS)
  - Cambridge Crystallography Data Centre (CCDC)
  - Royal Society of Crystallography
  - Chemistry Central
  - Reciprocal Net (US, Australia, UK)
- Open standards and software e.g. CIF, checkcif, CML, INChI
- Culture for sharing/depositing data (CCDC)
- Well-established workflow for crystallography experiments
- One dominant file format (CIF) - international exchange format
- Example: http://homes.ukoln.ac.uk/~lismp/IDCC2007/RINetCIF.htm



Partial view of an RI Network for the CIF File format

Capturing RI: Engineering Data

- Engineering is a broad area (mechanical, electrical, civil, architecture, construction, defence etc.)
- Vested commercial interests
- Proliferation of proprietary file formats
- Closed software solutions
- IGES 5.3: first popular exchange format (STEP still immature)
- Example: http://homes.ukoln.ac.uk/~lismp/IDCC2007/iges.html



Partial view of an RI Network for the IGES 5.3 File format

Capturing RI: Challenges and Issues (1)

- Constructing RI Networks is time-consuming and non-trivial
  - Sheer amount of information to be structured and documented
  - Take tacit, unstructured and dynamic knowledge and make it explicit with encoded relationships to enable automated processing (Semantic Web)
  - Domain expertise required for comprehensive and robust RI networks
    - Need simple, automated tools and procedures
    - Semantic Web (Web 3.0) technology based tools
  - Not clear when to end the recursion
    - Designated Community and associated Knowledge Base difficult to define
    - Designated Community and associated Knowledge Base are dynamic
  - Need robust search and retrieval of RI to build RI networks
- Continuous Monitoring to keep RI fit for purpose
  - Designated Community
  - Knowledge Base
  - maintenance of RI and RI networks

Capturing RI: Challenges and Issues (2)

- Classification of RI
  - In the OAIS is at a very high level (structure, semantic, other)
  - RRoRI has a more granular but generic classification
  - Will impact on search and retrieval of RI
  - Likely to need domain based classification to cater for
    - Domain or application specific RI (e.g. INChI, particular instrumentation)
    - Significant characteristics of specialist data (e.g. INChI)
- IPR and Rights
  - Easier in domains that use open standards and software (e.g. crystallography, although pharmaceuticals is a counter-example)
  - Computer Aided Design (CAD)
    - Intimate connection between models, formats and software
    - Formats are proprietary and unpublished
    - Format specifications may not be sufficient to interpret files (need software as well – proprietary and closed)

Capturing RI: Challenges and Issues (3)

- Technical Infrastructure
  - Need to record CPID as part of (preservation) metadata
  - Resolver service for CPID to enable automatic traversal of RI network
  - Continuous curation and maintenance of CPID, RI, RI Label and RI networks
  - Effective search and retrieval of RI
- Cost/Benefit/Risk Analysis
  - Curation and preservation are costly activities which require recurring, long-term funding commitments
  - RI underpins other strategies e.g. migration, emulation, normalisation
  - Cost/Benefit/Risk models will become more and more important
    - e.g. recently proposed model from the LIFE Project

      $L_t = Aq + I_t + M_t + Ac_t + S_t + P_t$

      (Cost = Aquisition + Ingest + Metadata + Access + Storage + Preservation)

Conclusions

- Need digital curation throughout the useful lifetime of digital data
  - Legal and safety requirements
  - Maximise potential of digital data
  - Maximise investment in digital data
- Plan from the outset for longevity and sustainable access
- A preservation strategy based on RI depends on a global, well-engineered, distributed infrastructure of RI
  - Needs coordination, collaboration and globally shared effort
  - Mining of RI networks for inference purposes
- Creation of robust RI networks requires domain expertise
- Likely to be gaps in global networks of RI
  - Business case for using a store of RI is clear, however the case for submitting RI to the global effort is less clear (commercial, IPR etc.)

Acknowledgements

- David Giaretta, Stephen Rankin, Brian McIlwrath (STFC, DCC, CASPAR)
- Simon Coles (NCS, eBank-UK)
- Chris McMahon (University of Bath, KIM)
- JISC
- EPSRC/ESRC

Further Information

- OAIS Reference Model: http://www.ccsds.org/documents/650x0b1.pdf
- DCC Development White Paper “DCC Approach to Digital Curation under Development”: http://dev.dcc.ac.uk/twiki/bin/view/Main/DCCApproachToCuration
- CASPAR Project: http://www.casparpreserves.eu
- M. Patel and S. Coles, "A Study of Curation and Preservation issues in the eCrystals Data Repository and proposed federation", Sept. 2007 http://www.ukoln.ac.uk/projects/ebank-uk/curation/
- eBank-UK Project http://www.ukoln.ac.uk/projects/ebank-uk/
- Knowledge & Information Management through Life: A Grand Challenge Project http://www-edc.eng.cam.ac.uk/kim/

Questions?

Thank you

Manjula Patel, Alexander Ball
UKOLN, University of Bath, UK
{m.patel, a.ball}@ukoln.ac.uk

http://www.ukoln.ac.uk/